Published online 26 November 2014 *Nucleic Acids Research, 2015, Vol. 43, Database issue* ***D825–D830***
doi: 10.1093/nar/gku1123

arrayMap 2014: an updated cancer genome resource

Haoyang Cai[1,2,3,*], Saumya Gupta[1,2], Prisni Rath[2,4], Ni Ai[1,2] and Michael Baudis[1,2,*]

[1]Institute of Molecular Life Sciences, University of Zurich, 8057 Zurich, Switzerland, [2]Swiss Institute of Bioinformatics, 8057 Zurich, Switzerland, [3]Center of Growth, Metabolism, and Aging, Key Laboratory of Bio-Resources and Eco-Environment, College of Life Sciences, Sichuan University, Chengdu 610064, Sichuan, China and [4]Centre for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland

Received September 1, 2014; Revised October 19, 2014; Accepted October 25, 2014

ABSTRACT

Somatic copy number aberrations (CNA) represent a mutation type encountered in the majority of cancer genomes. Here, we present the 2014 edition of arrayMap (http://www.arraymap.org), a publicly accessible collection of pre-processed oncogenomic array data sets and CNA profiles, representing a vast range of human malignancies. Since the initial release, we have enhanced this resource both in content and especially with regard to data mining support. The 2014 release of arrayMap contains more than 64 000 genomic array data sets, representing about 250 tumor diagnoses. Data sets included in arrayMap have been assembled from public repositories as well as additional resources, and integrated by applying custom processing pipelines. Online tools have been upgraded for a more flexible array data visualization, including options for processing user provided, non-public data sets. Data integration has been improved by mapping to multiple editions of the human reference genome, with the majority of the data now being available for the UCSC hg18 as well as GRCh37 versions. The large amount of tumor CNA data in arrayMap can be freely downloaded by users to promote data mining projects, and to explore special events such as chromothripsis-like genome patterns.

INTRODUCTION

Somatic genomic alterations refer to DNA sequence changes that are acquired during an individual's lifetime in the body's tissues (1,2). The type of unbalanced structural alterations which are called copy number aberrations/alterations (CNAs) are important subclasses of somatic DNA changes, including duplication, multi-copy amplification as well as homo- or heterozygous deletions of chromosomal segments (3). These frequently complex aberrations have been found in nearly all human tumor types, with regions spanning from several dozens of nucleotide bases to whole chromosomes (4–6). CNAs contribute to the initiation and progression of human malignancies by activating oncogenes, silencing tumor suppressor genes or disturbing gene expression through the involvement of regulatory elements (7,8). In the last two decades, array comparative genomic hybridization (aCGH) technologies have revolutionized cancer genome research by allowing the genome-wide detection of CNAs with high spatial resolution (9,10) (we use the term 'aCGH' both for dual color experiments as well as for single color oligonucleotide arrays that rely on external reference data sets).

The tens of thousands of tumor samples profiled by genomic arrays and deposited in public repositories allow researchers to identify patterns of non-random CNA events related to different cancer types, and to pinpoint involvement of specific cancer genes (6,11,12). A number of databases providing curated CNA data are available online, such as CaSNP (13), CanGEM (14) and Progenetix (15). These resources typically focus on particular data type, are derived from a restricted range of array platforms or do not contain probe-level data representation.

The public version of arrayMap was launched in 2012 (16) as a reference resource for array based genome data sets of copy number imbalances in human malignancies. It presents pre-processed cancer genome data, mainly derived from processed NCBI Gene Expression Omnibus (GEO) (17) and EBI ArrayExpress (18) data sets, but also including user provided and publication derived data, and provides online tools to perform basic data analysis and visualization. Users can freely download probe-level and segmented genomic array data from the web site. Typical uses of arrayMap data include investigation of potential markers for cancer diagnosis and therapy; identification of particular low incidence events (e.g. chromothripsis-like patterns) (19–21); large-scale data mining, such as construction of specific cancer type CNA patterns, and comparison of arrayMap data with users' pre-publication data sets. Here, we summarize new developments in arrayMap content and utilities, which aim to increase data coverage and accuracy and im-

*To whom correspondence should be addressed. Tel: +41 44 635 3486; Fax: +41 44 635 68 11; Email: michael.baudis@imls.uzh.ch
Correspondence may also be addressed to Haoyang Cai. Tel: +86 28 85418843; Fax: +86 28 85412571; Email: haoyang.cai@gmail.com

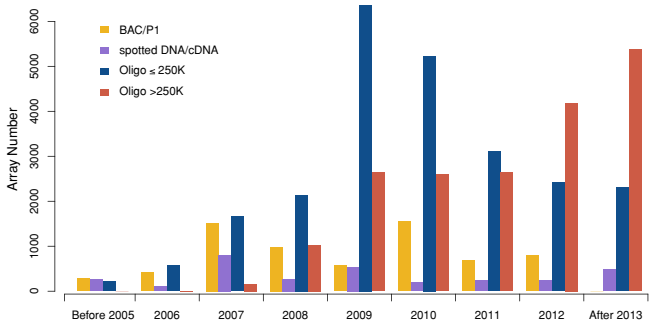

Figure 1. Distribution of arrays archived in arrayMap according to platform types and resolution. The usage of higher resolution platforms increased dramatically in recent years, with a concomitant decrease of especially low resolution BAC/P1 platforms.

portantly facilitate the use of this resource through a documented data interface.

DATA CONTENT UPDATES

Data growth

At the time of its launch, the original arrayMap edition contained about 40 000 arrays from 260 different platforms, representing 224 cancer diagnoses as defined in the International Classification of Diseases in Oncology (ICD-O 3) (22). For the 2014 edition, the absolute number of data sets has been increased to 64 814 genomic copy number arrays from 985 experimental series, involving 343 array platforms. The primary data had been published in more than 700 original publications, and now represents 252 ICD-O cancer entities (Table 1). Over time, relatively low resolution array platforms are replaced by higher resolution or multi-function platforms. At the moment, the platform with the highest probe numbers in arrayMap contains about 2.2 million individual probes. In line with this trend, ∼60% of the added arrays contain more than 250K probes. Since the data generated by high resolution arrays increased rapidly in recent years, we anticipate that this growth trend will continue (Figure 1) with special impact on the detection of focal genomic imbalances.

In this update, most novel samples were integrated from the NCBI GEO repository (17). Our main data selection criteria are that the data must be derived from human tumor samples and, where available, related germline DNA reference samples hybridized on single or dual color genomic array platforms. While primarily focusing on arrays with at least full autosomal coverage, we also integrated several studies with limited genome coverage which may provide useful information regarding gene specific CNAs in certain cancer types. In general, we used the formerly described pipeline (16) to re-process different data types. Briefly, for Affymetrix CEL files, we applied the aroma.affymetrix R package with the CRMAv.2 method (23) but utilized in-house scripts for data sets with available normalized probe intensity values. All probe signals were converted to log2 values, and Circular Binary Segmentation algorithm (24) was used for segmentation. For each array, empirical thresholds were assigned to call genomic gains and losses.

At the time of writing, data in arrayMap represent 252 ICD-O morphology codes. The largest of these are with 9551 samples 'adenocarcinoma, NOS' (8140/3; contains samples from e.g. prostate, gastric, colorectal and lung adenocarcinomas) and with 8188 samples 'invasive carcinoma of no special type' (8500/3; default histology for the majority of breast cancer samples; Figure 2). On the other end, 25 histologies are represented through a single array, among them e.g. 'giant cell sarcoma' and 'islet cell carcinoma'. The complete list of ICD-O histologies is available through the supplements or can be accessed through the data API (application program interface; see below) at http://arraymap.org/api/?db=arraymap&api_out=icdmlist&icdm_m=0,8,9. Among the clinical entities, breast cancers constitute by far the largest category (8837 arrays) followed by non-small cell lung carcinomas (4112 arrays), acute myeloid leukemias (3641 arrays) and colorectal carcinomas (3047 arrays; Supplementary Table 2). The complete list is provided as supplement, or can be generated through calling 'http://arraymap.org/api/?db=arraymap&api_out=cgrouplist&icdm_m=0,8,9'.

Compared to managed large-scale projects with frequent focus on a few predominant cancer types, the assembly of data from hundreds of individual studies has an inherent advantage in representing the heterogeneous landscape of human malignancies. As an example, when matching the arrayMap data to the content of the TCGA / 'Pan-cancer project' (25,26), one can observe that the 12 tumor types on which the 'Pan-Cancer' study has focused so far correspond to about half of the sample content in arrayMap (Table 2). While the efforts of the leading TCGA and ICGC (27) projects aim at a detailed multi-level description of molecular aberrations and their biological impact on cancer progression, the proportion of arrayMap samples from cancers not represented in those studies should serve as a reminder of the large number of 'rare' tumor types encountered in oncological practice, and the gap in our understanding of their molecular mechanisms. In our opinion, the arrayMap resource can prove especially useful in promoting oncogenomic data mining projects aimed at identifying exceptional tumor biologies.

Besides the focus on cancer samples, the new edition of arrayMap also contains normal tissue samples that were used as controls in cancer profiling experiments. The amount of high resolution data from more than 8000 normal samples now allows for the creation of a matched number variation track, without relying on external resources (28–30) (Supplementary Figure S1). These data can be used to perform robust CNA data analysis, e.g. through providing a veto filter for the evaluation of focal (< 3–5 Mb) CNA events, which usually cannot be distinguished from germline variations without matched non-tumor samples.

Genome reference assembly mapping

In the first release of arrayMap, all genomic mapping information for probe positions and derived CNA segments was converted to the human genome assembly UCSC hg18 (NCBI Build 36.1) (31, http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/data/), to allow for the integration of the different platform types and exper-

Table 1. arrayMap content increase (as of 31 August 2014) compared to the initial release

Database content	Number of entries 2012	Number of entries 2014
Arrays	~ 40 000	64 814
Series	533	985
Platforms	260	343
ICD-O cancer types	224	252
Publications	638	716
Patients	15 634	23 713

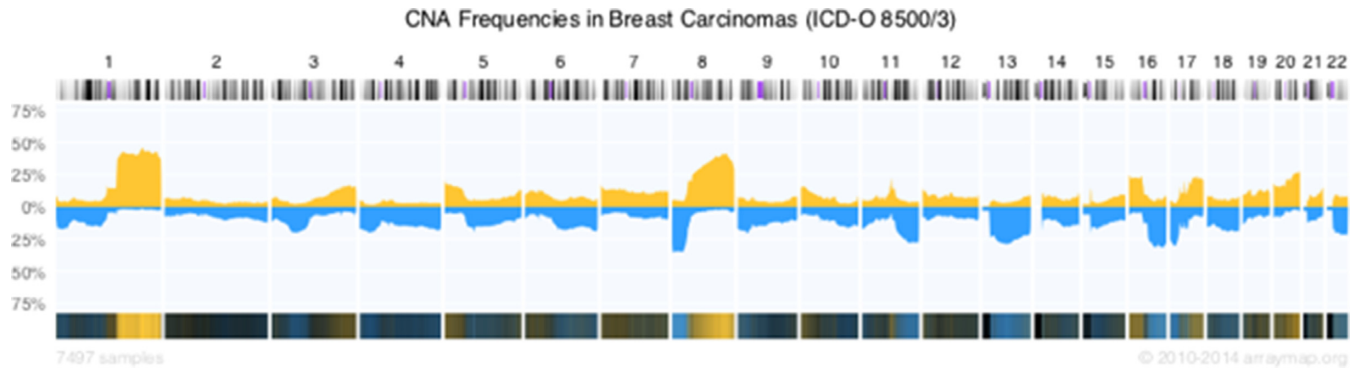

Figure 2. Frequency of copy number gains (up, yellow) and losses (down, blue), derived from 7497 breast cancer arrays (here limited to ICD 8500/3), as represented in the arrayMap database. For each of the included arrays, experiment specific CNA data can be accessed individually.

Table 2. The proportions of TCGA 'Pan-Cancer' tumor diagnoses in arrayMap

Tumor types	Array numbers	Percentage
Breast carcinoma	8837	15.8%
Lung adenocarcinoma	3464	6.2%
Colon adenocarcinoma	3085	5.5%
Lymphoblastic acute myeloid leukemia	2848	5.1%
Glioblastoma	2517	4.5%
Ovarian carcinoma	2477	4.4%
Kidney renal clear-cell carcinoma	1473	2.6%
Head and neck squamous carcinoma	1423	2.5%
Uterine cervical and endometrial carcinoma	1013	1.8%
Lung squamous carcinoma	839	1.5%
Bladder carcinoma	689	1.2%
Rectal adenocarcinoma	250	0.4%
Other tumor types	27 101	48.5%

imental results. For this goal, a pipeline was generated to map the genomic positions for the thousands of array probes to the common 'Golden Path' edition. In recent years, new genome assemblies have been provided, (UCSC hg19 / GRCh37 and recently UCSC hg20 / GRCh38) with GRCh37 now frequently being used for referencing genomic array coordinates. When updating data from hg18 to newer assemblies, the change of probe coordinates may affect the composition of previously called CNA regions through un-mapping of some coordinates. To minimize this problem, for arrays with available probe values we first remapped all probe positions to GRCh37 using the UCSC Genome Browser's liftOver tool with intermediate BED files (30), and then re-segmented based on the derived probe positions. Although a few probes failed to be remapped during this procedure, the average remapping rate was as high as 99%. For a subset of e.g. literature derived data sets, segmentation data were processed directly. At the moment we are planning to migrate the database to the newest GRCh38 assembly.

NEW AND ONGOING DEVELOPMENTS

Web front end and data visualization

Some of the main strengths of the arraymap repository are the pre-computed visualization of some 10 000 probe-level genomic array data sets, as well as the graphical representation of CNA distributions based on curated clinical information, most notably the samples' assignment to standardized diagnostic categories based on the WHO's ICD-O 3 schema (22). Since the arrayMap resource is based on the software framework developed for the Progenetix project (32), the data search and visualization updates reported in the 2014 Progenetix update (15) apply for the arrayMap resource, too. For the data selection, these include predefined aggregate data for ICD entities, tumor loci, SEER (33, http://www.seer.cancer.gov/popdata) categories as well as 'clinical groups', referring to samples with a common clinical context (e.g. 'carcinomas: breast carcinomas' including all types of epithelial breast tumors). Another option introduced with the latest Progenetix update and now applied to

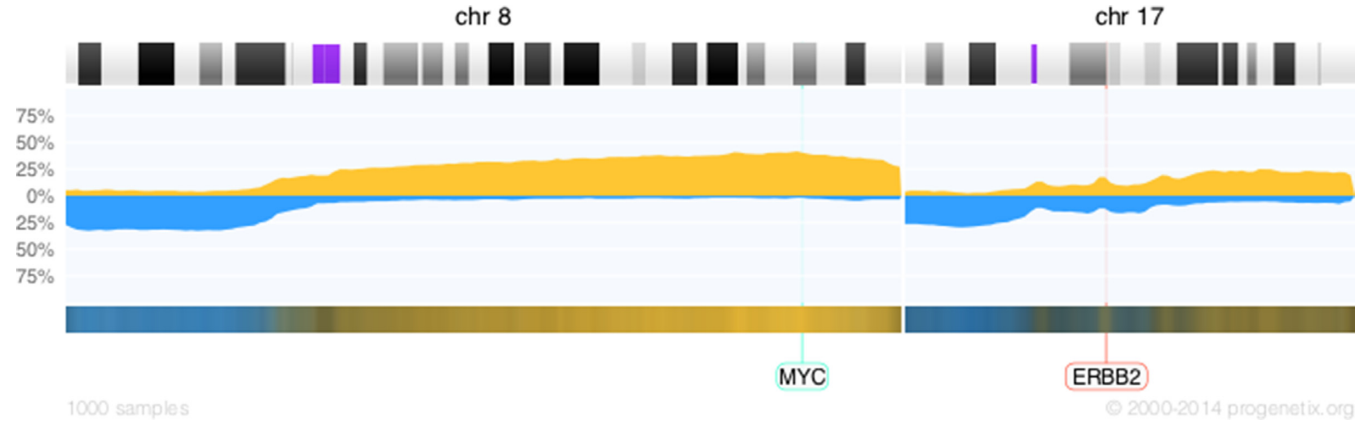

Figure 3. The histogram profile of breast cancer samples focusing on chromosomes 8 and 17 generated by the API, displaying the frequency of copy number gains (up, yellow) and losses (down, blue). Genomic map positions of two genes were included in the API call, and therefore are labeled at the bottom of the figure.

arrayMap is the geographic mapping of the included studies according. In the case of arrayMap, samples are mapped based on the submitting information from GEO, with a fallback to the corresponding author of the related publication. While this feature is not as useful as e.g. patient data derived origin mapping, it nevertheless offers a fast overview about enters with research activity in the corresponding cancer types and may support networking activities between research groups. Although the mapping information does not disclose the samples' origins, the almost complete lack of data sets for large swaths of the globe (e.g. Africa, central Asia, South America) points to unmined cancer genome resources and paucity of research into possible epidemiological and environmental factors.

API

The 2014 arrayMap release is the first to provide a RESTful data API. The API provides a variety of query and output parameters, with URL formatted (GET) requests returning server side processed data as JSON (Java Script object notation) objects, test/tabular data or images suitable for direct embedding or storage. A detailed and continuously updated documentation can be found online, in the arrayMap/Progenetix user guide at http://wiki.progenetix.org.

API Example 1: Sample data as JSON

The following query will return all samples from ICD-O 3 codes starting with '817' (i.e. hepatocellular adenomas/carcinomas) from the arrayMap collection:

http://arraymap.org/api/?db=arraymap&api_out=samples&api_doctype=json&icdm_m=817

API Example 2:

The query will return a gain/loss frequency histogram for chromosomes 8 and 17, derived from 1000 random samples of ICD-O 8500/3 (breast carcinoma - 'invasive carcinoma of no special type'), in the form of a PNG data stream (Figure 3):

http://arraymap.org/api/?db=arraymap&markers_m=[MYC]8:128816862-128822853,[ERBB2]17:35104766-35138441&icdm_m=8500/3&randno=1000&api_out=histogram&chr2plot=8,17

API Example 3:

The query will return the number of samples in arrayMap for ICD-O 8500/3, which have gain CNAs overlapping both the MYC and ERBB2 loci:

http://arraymap.org/api/?db=arraymap&locus_m=8:128816862-128822853:1,17:35104766-35138441:1&icdm_m=8500/3&;api_out=count

API: R

With the ability to access the status matrix directly, one easily can import the data into an R data frame:

pgframe <- read.table(url('http://arraymap.org/api/?icdm_m=814&db=arraymap&api_out=matrix'), header = T, sep = '\t', na = 'NA')

For the segment file, the same applies with 'output=segments':

segtable <- read.table(url('http://arraymap.org/api/?text_m=sezary&db=arraymap&api_out=segments'), header = T, sep = '\t', na = 'NA')

To facilitate R integration of Progenetix/arrayMap data, we have developed a simple access function 'pgDataLoader' which can currently be accessed through GitHub (https://github.com/progenetix/pgRpi/). This publication's supplements include an example use case, describing the generation of gene specific Kaplan-Meyer survival plots from arrayMap data.

User managed data

In this version of arrayMap, we provide some online support for the analysis and visualization of user private (i.e. pre-publication) array data sets. After registration by email, users are able to use on site storage facilities and recall previous performed analyses. For example, users can directly upload and visualize segmentation files, sample tables with ISCN karyotypes, or JSON files from a previous analysis. Data subsets from database queries can be reloaded and used for filtering and replotting. Additionally to these options, the analysis of raw / pre-processed probe data sets is supported in collaborative projects. Analysis input here can be e.g. Affymetrix genotyping array raw data (.CEL files),

other platforms from log2 value lists, and pre-existing segmentation data.

CONCLUSIONS AND FUTURE PERSPECTIVES

arrayMap is developed to provide a one stop resource of genomic copy number profiles of human tumors, as well as a series of online tools for meta-data analysis and mining. Although arrayMap is tightly integrated with and shows some content overlap with the Progenetix resource (http://www.progenetix.org), both data collections offer different scopes and data paradigms (Supplementary Figure S4). In contrast to arrayMap, which displays pre-processed but loosely evaluated experimental array data, Progenetix annotations are based on sample specific copy number data, from different technologies (chromosomal CGH, genomic arrays, genome sequencing), were the 'called' CNA had been either provided through a publication, or had been assessed from an active evaluation of the original experimental data. While the Progenetix resource has an advantage in providing genomic aberration data for an even wider diagnostic range than arrayMap (362 versus 252 ICD-O entities), it is more heterogeneous with respect to included technologies and spatial resolution of the CNA data sets (e.g. cytoband based cCGH data) which limits e.g. the detection of rare focal CNA events.

Since the launch of the resource in 2012, arrayMap underwent a number of quantitative, qualitative and functional improvements, most notably the increase in included data sets and scope of represented cancer entities, as well as the addition of programmatic access methods and Progenetix based selection and visualization updates. For the future expansion of the arrayMap resource, we are evaluating the additional inclusion of data sets from multi-functional platforms (e.g. methylation arrays, mutation-specific probe sets). Moreover, a robust platform agnostic quality rating system is under development, and will be integrated in our database. For the overall data set expansion, we intend to follow an incremental, dynamic update policy, with bi-annual reassessments of major data content and feature changes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Haoyang Cai (China Scholarship Council, University of Zurich URPP "Systems Biology"); Prisni Rath (Swiss Institute of Bioinformatics); Romanian-Swiss Research Programme [RSRP IZERZ0-142305]. Funding for open access charge: university funds.
Conflict of interest statement. None declared.

REFERENCES

1. Albertson,D.G., Collins,C., McCormick,F. and Gray,J.W. (2003) Chromosome aberrations in solid tumors. *Nat. Genet.*, **34**, 369–376.
2. Stratton,M.R., Campbell,P.J. and Futreal,P.A. (2009) The cancer genome. *Nature*, **458**, 719–724.
3. Greenman,C., Stephens,P., Smith,R., Dalgliesh,G.L., Hunter,C., Bignell,G., Davies,H., Teague,J., Butler,A., Stevens,C. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153–158.
4. Beroukhim,R., Mermel,C.H., Porter,D., Wei,G., Raychaudhuri,S., Donovan,J., Barretina,J., Boehm,J.S., Dobson,J., Urashima,M. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.
5. Kim,T.-M., Xi,R., Luquette,L.J., Park,R.W., Johnson,M.D. and Park,P.J. (2013) Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Res.*, **23**, 217–227.
6. Baudis,M. (2402007) Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC Cancer*, **7**, 226.
7. Radtke,F. and Raj,K. (2003) The role of notch in tumorigenesis: oncogene or tumour suppressor? *Nat. Rev. Cancer*, **3**, 756–767.
8. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–D950.
9. Solinas-Toldo,S., Lampel,S., Stilgenbauer,S., Nickolenko,J., Benner,A., Döhner,H., Cremer,T. and Lichter,P. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.
10. Pinkel,D., Segraves,R., Sudar,D., Clark,S., Poole,I., Kowbel,D., Collins,C., Kuo,W.-L., Chen,C., Zhai,Y. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.
11. Zack,T.I., Schumacher,S.E., Carter,S.L., Cherniack,A.D., Saksena,G., Tabak,B., Lawrence,M.S., Zhang,C.-Z., Wala,J., Mermel,C.H. *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, **45**, 1134–1140.
12. Kumar,N., Cai,H., Mering,C.V. and Baudis,M. (2012) Specific genomic regions are differentially affected by copy number alterations across distinct cancer types, in aggregated cytogenetic data. *PLoS ONE*, **7**, e43689.
13. Cao,Q., Zhou,M., Wang,X., Meyer,C.A., Zhang,Y., Chen,Z., Li,C. and Liu,X.S. (2011) CaSNP: a database for interrogating copy number alterations of cancer genome from SNP array data. *Nucleic Acids Res.*, **39**, D968–D974.
14. Scheinin,I., Myllykangas,S., Borze,I., Bohling,T., Knuutila,S. and Saharinen,J. (2007) CanGEM: mining gene copy number changes in cancer. *Nucleic Acids Res.*, **36**, D830–D835.
15. Cai,H., Kumar,N., Ai,N., Gupta,S., Rath,P. and Baudis,M. (2014) Progenetix: 12 years of oncogenomic data curation. *Nucleic Acids Res.*, **42**, D1055–D1062.
16. Cai,H., Kumar,N. and Baudis,M. (2012) arrayMap: a reference resource for genomic copy number imbalances in human malignancies. *PLoS ONE*, **7**, e36944.
17. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
18. Petryszak,R., Burdett,T., Fiorelli,B., Fonseca,N.A., Gonzalez-Porta,M., Hastings,E., Huber,W., Jupp,S., Keays,M., Kryvych,N. *et al.* (2014) Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42**, D926–D932.
19. Stephens,P.J., Greenman,C.D., Fu,B., Yang,F., Bignell,G.R., Mudie,L.J., Pleasance,E.D., Lau,K.W., Beare,D., Stebbings,L.A. *et al.* (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, **144**, 27–40.
20. Rausch,T., Jones,D.T.W., Zapatka,M., Stütz,A.M., Zichner,T., Weischenfeldt,J., Jäger,N., Remke,M., Shih,D., Northcott,P.A. *et al.* (2012) Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell*, **148**, 59–71.
21. Cai,H., Kumar,N., Bagheri,H.C., von Mering,C., Robinson,M.D. and Baudis,M. (2014) Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens. *BMC Genomics*, **15**, 1–13.
22. Fritz,A., Percy,C., Jack,A., Shanmugaratnam,K., Sobin,L., Parkin,D.M. and Whelan,S. (2000) *International Classification of*

Diseases for Oncology (ICD-O). 3rd edn. World Health Organization, Geneva.
23. Bengtsson,H., Wirapati,P. and Speed,T.P. (2009) A single-array preprocessing method for estimating fullresolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics*, **25**, 2149–2156.
24. Olshen,A.B., Venkatraman,E.S., Lucito,R. and Wigler,M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
25. Mclendon,R., Friedman,A., Bigner,D., Meir,E.G.V., Brat,D.J., Mastrogianakis,G.M., Olson,J.J., Mikkelsen,T., Lehman,N., Aldape,K. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
26. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
27. International Cancer Genome Consortium, Hudson,T.J., Anderson,W., Artez,A., Barker,A.D., Bell,C., Bernabé,R.R., Bhan,M.K., Calvo,F., Eerola,I. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
28. Redon,R., Ishikawa,S., Fitch,K.R., Feuk,L., Perry,G.H., Andrews,T.D., Fiegler,H., Shapero,M.H., Carson,A.R., Chen,W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
29. Lafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
30. Karolchik,D., Barber,G.P., Casper,J., Clawson,H., Cline,M.S., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, **42**, D764–D770.
31. Church,D.M., Schneider,V.A., Graves,T., Auger,K., Cunningham,F., Bouk,N., Chen,H.-C., Agarwala,R., Mclaren,W.M., Ritchie,G.R. *et al.* (2011) Modernizing reference genome assemblies. *PLoS Biol.*, **9**, e1001091.
32. Baudis,M. and Cleary,M.L. (2001) Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics*, **17**, 1228–1229.
33. Surveillance, Epidemiology, and End Results (SEER) Program Populations (1969–2012) *National Cancer Institute, DCCPS, Surveillance Research Program*. Surveillance Systems Branch, released March 2014.